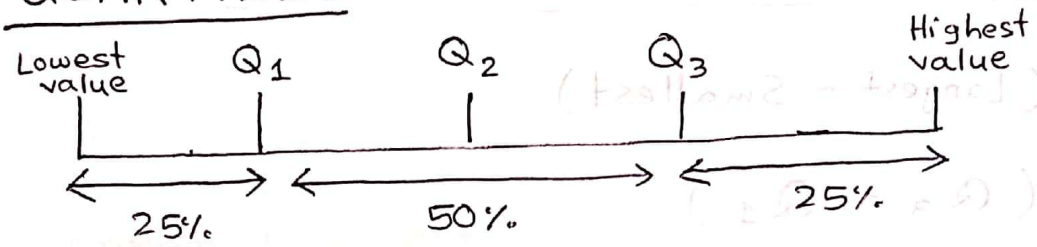
2 MEASURES OF LOCATION AND SPREAD

QUARTILES



Q_1 – Lower quartile $= \frac{n}{4}$ th data value

Q_3 – Upper quartile $= \frac{3n}{4}$ th data value

Q_2 – Median $= \frac{n}{2}$ th data value

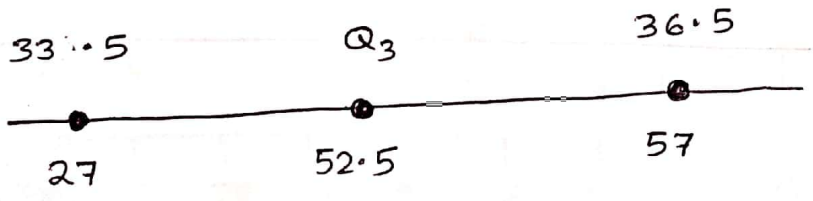
INTERPOLATION

When interpolation is used, it is assumed that the data values are evenly distributed within each class.

Example –

Time spent	30 - 31	32 - 33	34 - 36	37 - 39
Frequency	2	25	30	13

Upper quartile: $\frac{3}{4} \times (2+25+30+13) = 52.5^{th}$ value



$$\frac{Q_3 - 33.5}{36.5 - 33.5} = \frac{52.5 - 27}{57 - 27}$$

$$\Rightarrow Q_3 = \frac{25.5}{30} \times 3 + 33.5$$

$$= 36.05$$

Avishek

Scanned with CamScanner

Measures of Spread / Dispersion / Variation :

- Range (Largest - Smallest)
- IQR ($Q_3 - Q_1$)
- Interpercentile range

VARIANCE & STANDARD VARIATION

Variance : Another way to work out spread of data set

$$\text{Variance } (\sigma^2) = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{S_{xx}}{n}$$

$\bar{x}$ = mean

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} \quad [\text{summary statistic}]$$

$$\text{Standard deviation } (\sigma) = \sqrt{\text{variance}}$$

CODING

$$y = \frac{x - a}{b}$$

mean : $\bar{y} = \frac{\bar{x} - a}{b}$

standard deviation : $\sigma_y = \frac{\sigma_x}{b}$

Avishek

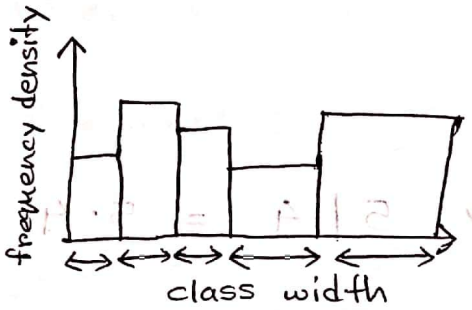
Scanned with CamScanner

3 REPRESENTATIONS OF DATA

Grouped continuous data is represented in histogram

area of bar = k × frequency $\quad A = F.k$

frequency density = $\frac{\text{frequency}}{\text{class width}}$ $\quad F.d = \frac{F}{c.w}$



Avishek

OUTLIERS Extreme value that lies outside the overall pattern of the data.

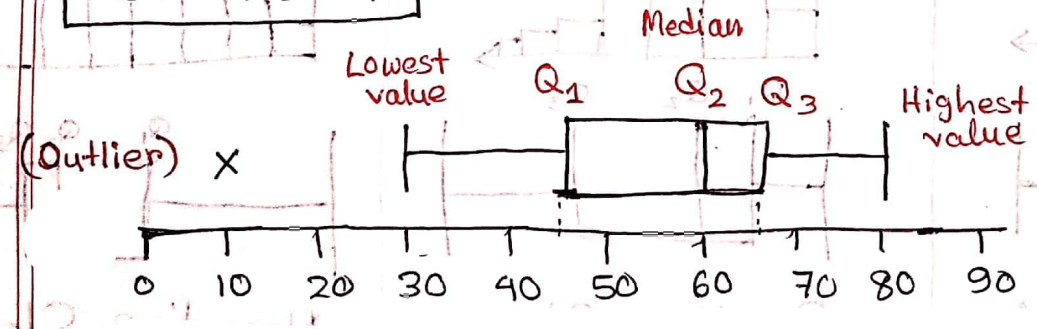
$Q_3 + k(Q_3 - Q_1)$

$Q_1 - k(Q_3 - Q_1)$

or, $1.5 \times IQR$

$1.5 \times (Q_3 - Q_1)$

BOX PLOT



Scanned with CamScanner

STEM & LEAF DIAGRAM

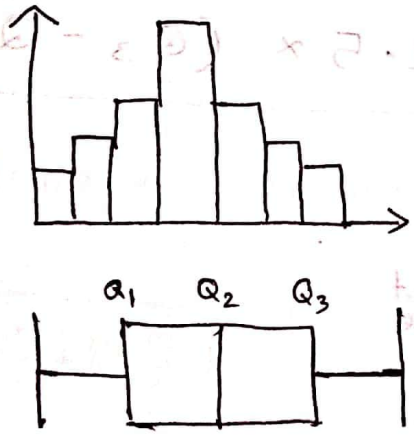
1·4	3·5	2·2	5·6	7·8	8·4	8·6	1·1	1·3
2·1	3·4	3·6	4·1	8·8	8·8	8·8	7·9	7·8

Stem	Leaf
1	1 3 4
2	1 2
3	4 5 6
4	1
5	6
6	
7	8 8 9
8	4 6 8 8 8

Avishek

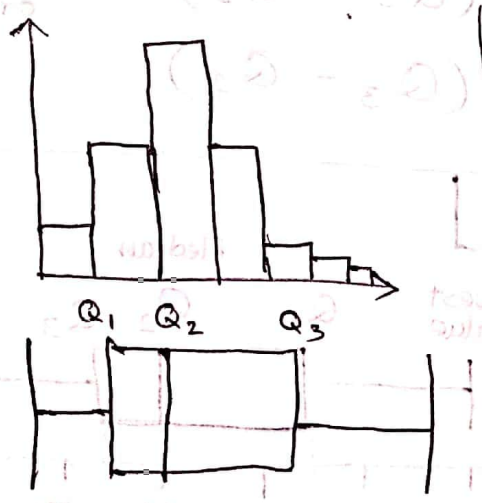
Key: 5 | 4 = 5·4

SKEWNESS



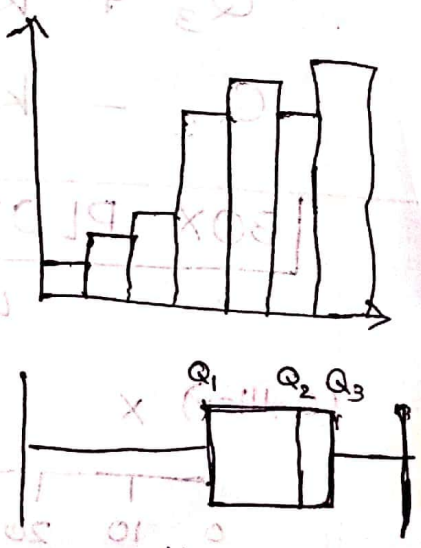
Symmetrical

$Q_2 - Q_1 = Q_3 - Q_2$



Positive Skew

$Q_2 - Q_1 < Q_3 - Q_2$



Negative Skew

$Q_2 - Q_1 > Q_3 - Q_2$

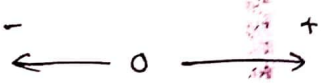
Scanned with CamScanner

Avishek

Data which are mostly at lower values have (+) skew.
" " " " " higher " " (-) " .

Mode = median = mean → symmetrical
Mode < median < mean → positive skew
Mode > median > mean → negative skew

Skewness : $\frac{3\,(\text{mean} - \text{median})}{\text{standard deviation}}$



COMPARING DATA

IQR - median comparison ✓ [If extreme values]
Mean - standard deviation comparison ✓

Removing anomalies is called cleaning the data.

Box plots can be used to compare two sets of data.

Scanned with CamScanner

Box plot comparison:

IQR, Range, Median

Histogram: (For continuous data)

$$F.d. = \frac{f}{c.w.}$$

$$f = kA$$

@Mindful Maths

Skew:

$Q_3 - Q_2 > Q_2 - Q_1$

Positive skew
(more values at start)

$Q_2 - Q_1 > Q_3 - Q_2$

Negative skew

No skew

$Q_3 - Q_2 = Q_2 - Q_1$

Skewness:

$$\frac{3\,(\text{mean} - \text{median})}{\text{standard deviation}}$$

Scanned with CamScanner

MEDIAN

☆ Position of median = $\frac{n+1}{2}$ ← List of numbers

✱ Position of median = $\frac{n}{2}$ ← Grouped data

☆ Median = $\frac{\text{Sum of middle no.}}{2}$ ← Median (if position in [List] d.p.)

✱ Median = Interpolation ← Grouped data (if position in d.p.)

LINEAR INTERPOLATION

Height (m)	Frequency
$0.55 \leq h \leq 0.6$	55
$0.6 \leq h < 0.65$	45
$0.65 \leq h < 0.7$	30
$0.7 \leq h < 0.75$	15
$0.75 \leq h < 0.8$	5

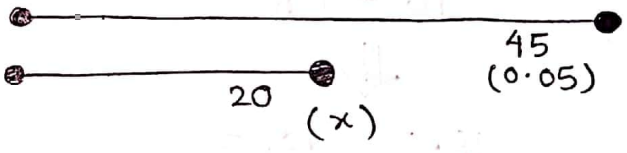
55 | 100

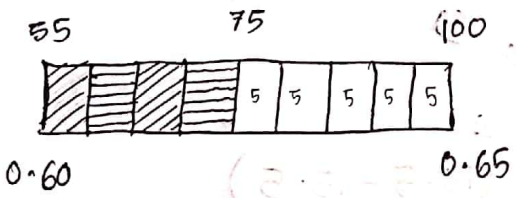
$n = 150$

median = $\frac{150}{2} = 75$ $(0.6 \leq h < 0.65)$

range of frequency: $55 < n < 100$

55^{th}	75^{th}	100^{th}
0.6	Median	0.65





$\frac{20}{45} = \frac{x}{0.05}$

$\Rightarrow x = \frac{1}{45}$

$0.6 + \frac{1}{45} = \frac{28}{45} = 0.62\dot{2}$

$\frac{x}{0.65-0.60} = \frac{75-55}{100-55}$

Median (h) = $0.62\dot{2}$

Make a cumulative frequency table for easier calculation

@Mindful Maths

Scanned with CamScanner

Weight (kg)	Frequency
10 - 12	7
13 - 15	2
16 - 18	9
19 - 20	4

Data table with GAPS between intervals.

← 11th term in this range

10 - 12

⇓

lower bound → 9.5 - 12.5 ← upper bound

↑

Class width = 3

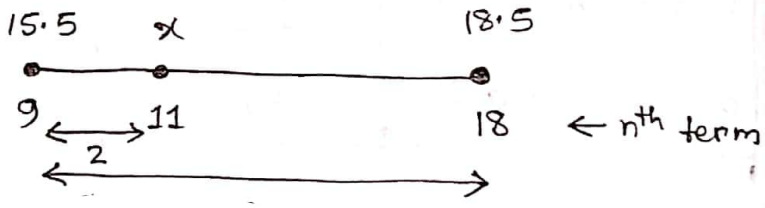
@Mindful Maths

Median: $\frac{\Sigma f}{2} = \frac{22}{2} = 11^{th}$ term

16 - 18

⇓

15.5 - 18.5



Class width = 3

$\frac{2}{9} \times (18.5 - 15.5)$

$= \frac{2}{9} \times 3$

$= \frac{2}{3}$

11^{th} term $= 15.5 + \frac{2}{3} = 16.16\dot{}$

Scanned with CamScanner

PERCENTILE

For percentile maths, always group data given, so NEVER ROUND OFF!

Example:

$n = 43$ What's the 57^{th} percentile? (P_{57})

ANS: $43 \times \frac{57}{100} = 24.51^{th}$

Inter Quartile Range

$$IQR = Q_3 - Q_1$$

Q_2 = median

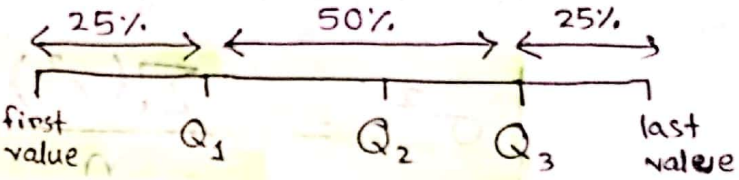
QUARTILES

Listed data:

$$Q_1 = \frac{1}{4} \times n$$

$$Q_2 = \frac{1}{2} \times n$$

$$Q_3 = \frac{3}{4} \times n$$



Grouped data:

Same formula as above BUT USE LINEAR INTERPOLATION

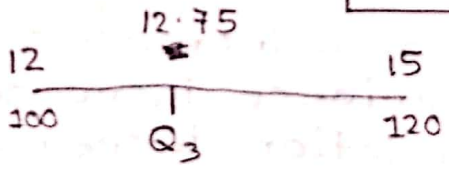
Do not round this value

Example:

$Q_3 = \frac{7+5+3+2}{4} \times 3$

$= 12.75^{th}$

Weight	Frequency
$80 \leq q < 90$	7
$90 \leq q < 100$	5
$100 \leq q < 120$	3
$120 \leq q < 200$	2



$$\frac{12.75 - 12}{15 - 12} = \frac{Q_3 - 100}{120 - 100}$$

$$\Rightarrow Q_3 = \frac{0.75}{3} \times 20 + 100$$

$$= 105$$

@Mindful Maths

Scanned with CamScanner

VARIANCE

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

$\bar{x}$ = mean
n = total frequency
$\sum$ = sum of
σ^2 = variance

Variance – Avg. squared distance from mean.

"Mean of squares minus square of mean"

msmsm

$$\sigma^2 = \frac{\sum (x^2)}{n} - (\bar{x})^2$$

STANDARD DEVIATION

$$\sigma = \sqrt{\text{variance}}$$

@Mindful Maths

$$\sigma^2 = \frac{\sum f x^2}{\sum f} - (\bar{x})^2$$

- If data increases by x times,
  variance increases by x^2 times
  s. deviation increases by x times

Scanned with CamScanner

CODING	Effect on $\bar{x}$	Effect on σ
$y = x + 10$	$\bar{x} + 10$	+ or − No effect
$y = 3x$	$\bar{x} \times 3$	$\sigma \times 3$
$y = 2x - 5$	$2 \times \bar{x} - 5$	-5 no effect $\sigma \times 2$

@Mindful Maths

Old mean $\bar{x}$	Old σ_x	Coding	New mean $\bar{y}$	New σ_y
36	4	$y = x - 20$	16	4
36	8	$y = 2x$	72	16
35	4	$y = 3x - 20$	85	12
40	3	$y = \frac{x}{2}$	20	$\frac{3}{2}$
11	27	$y = \frac{x+10}{3}$	7	9
300	25	$y = \frac{x-100}{5}$	40	5

$$\Downarrow$$

$$y = \frac{x}{5} - 20$$

Scanned with CamScanner

PROBABILITY

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	2	3	4	5	6

$$P(5) = \frac{4}{6 \times 6} = \frac{1}{9}$$

$$P(8 < x) = \frac{10}{36} = \frac{5}{18}$$

$\cap$ - intersection

$\cup$ - union

A' - complement

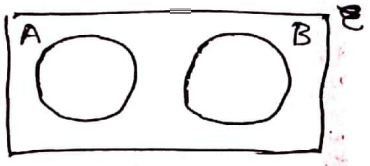
$\mathcal{E}$ - sample space

$\varnothing$ - null set

$\subset$ - subset

Avishek

When events have no outcomes in common, they are called <u>mutually exclusive</u>.



$$P(A \text{ or } B) = P(A) + P(B)$$

For independent events,

$$P(A \text{ and } B) = P(A) \times P(B)$$

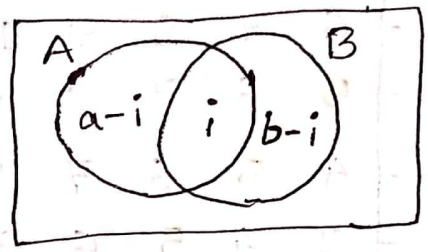
Probability that B occurs given that A already occured $= P(B \mid A)$

$$P(A \mid B) = P(A \mid B') = P(A)$$ [Independent events]

Scanned with CamScanner

$P(A \cup B) = a + b - i$

$P(A \cap B) = i$



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

Avishek

Scanned with CamScanner

PROBABILITY

1. Basic probability (Dice, toss etc)
2. Venn Diagrams
3. Mutually Exclusive / Independent Events
4. Tree Diagram

SAMPLE SPACE DIAGRAM

Spinner 2 \ Spinner 1	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

$$P(5) = \frac{4}{4 \times 4} = 0.25$$

$$P(>5) = \frac{6}{16} = \frac{3}{8}$$

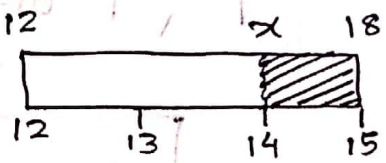
@ mindful maths

Sample space - Set of all possible outcomes.

Mass	$9 \leq h < 11$	$11 \leq h < 12$	$12 \leq h < 15$
frequency	9	3	6

$$P(<12) = \frac{9+3}{9+3+6} = \frac{2}{3}$$

$$P(\geq 14) = \frac{4}{18} = \frac{2}{9}$$

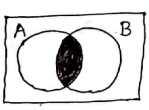


$$x = \frac{2}{3} \times 6$$

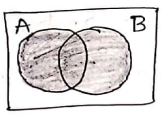
$$= 4$$

Scanned with CamScanner

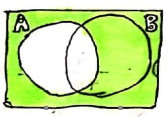
VENN DIAGRAMS



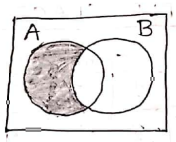
- A and B
- Intersection
- $A \cap B$



- A or B
- union of A & B
- $A \cup B$



- not A
- ~~union~~
- A'

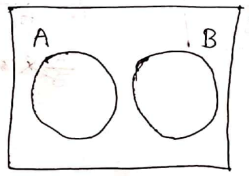


- A and not B
- $A \cap B'$

Mutually Exclusive

$P(A \text{ and } B) = P(A \cap B) = 0$

$P(A \text{ or } B) = P(A) + P(B)$



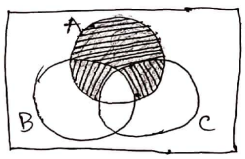
← no intersection, no common, exclusive

INDEPENDENT EVENTS

If A and B are independent:

$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B)$

$(A \cup B \cup C)' = A' \cap B' \cap C'$



$A \cap (B \cap C)'$

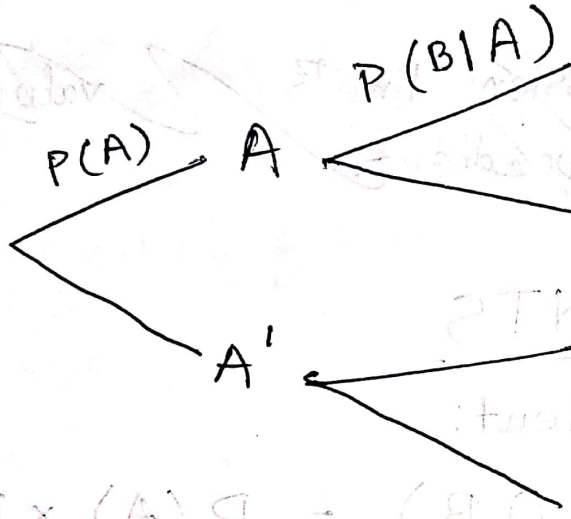
intersection = $\cap$ = and = multiplication

union = $\cup$ = or = addition

@ mindful maths

Scanned with CamScanner

CONDITIONAL PROBABILITY



$P(B \mid A)$

"given that"

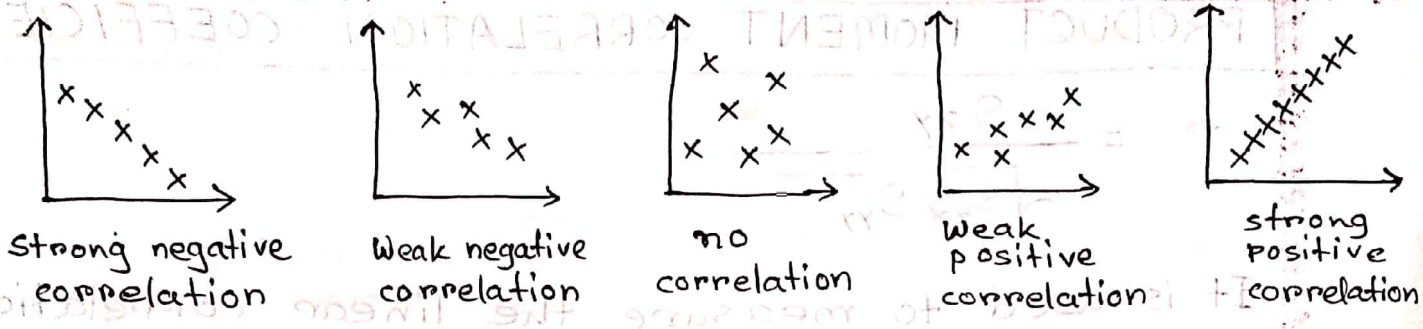
"B occured given that A occured"

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

@ mindful maths

Scanned with CamScanner

5 CORRELATION AND REGRESSION

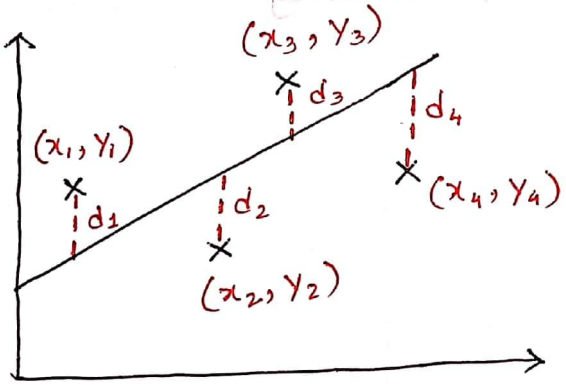


LINEAR REGRESSION

Avishek

Least Squares Regression Line is useful in statistics and one of the type of line of best fit.

This is the straight line that minimises the sum of the squares of the vertical distances of each data point from the line.



$$\sum (\delta_i^2)$$

Regression line of y on x : $y = a + bx$

$$b = \frac{S_{xy}}{S_{xx}}, \quad a = \bar{y} - b\bar{x}$$

(i) Interpolation and Extrapolation

(ii) Regression line is a valid model when data shows linear correlation.

(iii)
$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

Scanned with CamScanner

PRODUCT MOMENT CORRELATION COEFFICIENT

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

It is used to measure the linear correlation between two variables.. (-1 to 1)

Linear coding of data set x_i

$$P_i = a x_i + b$$

$$P_i = \frac{x_i - a}{b}$$

Avishek

r is not affected by coding

Scanned with CamScanner

CORRELATION

Stronger the (linear) correlation, the more suitable a linear regression line is.

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$\sigma^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$$

$$\therefore \quad \sigma^2 = \frac{S_{xx}}{n}$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$

Variance $(x) = \frac{S_{xx}}{n}$

Covariance $(x, y) = \frac{S_{xy}}{n}$

@ mindful maths

Product Moment Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$r = 1$ (strong +ve correlation)
$r = 0$ (no correlation)
$r = -1$ (strong -ve correlation)

Example:

Man	A	B	C	D	E	F
Circumference (x)	31.1	33.3	30.0	31.5	35.0	30.2
Height (y)	36	37	38	38	40	40

$\sum x = 191.1$
$\sum y = 229$
$\bar{x} = 31.85$
$\bar{y} = 38.17$
$\sum x^2 = 6105.39$
$\sum y^2 = 8753$
$\sum xy = 7296.7$
$n = 6$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 6105.99 - \frac{191.1^2}{6} = 18.855$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 8753 - \frac{229^2}{6} = 12.833$$

$$S_{xy} = \sum xy - \frac{\sum x \cdot \sum y}{n} = 7296.7 - \frac{191.1 \times 229}{6} = 3.05$$

$$r = \frac{S_{xy}}{\sqrt{S_{yy} S_{xx}}} = \frac{3.05}{\sqrt{12.833 \times 18.855}} = 0.196$$

Weak positive correlation

Scanned with CamScanner

If x increases k times and y increases q times, then

Variance of x : k^2 times bigger

S_{xx} : k^2 times bigger

S_{yy} : q^2 times bigger

S_{yx} : kq times bigger

r : unaffected

Line of best-fit:

$$y = a + bx$$

Mindful Maths

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \bar{y} - b\bar{x}$$

Scanned with CamScanner

DISCRETE RANDOM VARIABLE

$$P(x) = \begin{cases} 0.05x & x = 1, 3 \\ 0.1x & x = 2, 4 \\ 0.2 & x = 5 \end{cases}$$

x	1	2	3	4	5
$P(x)$	0.05	0.2	0.15	0.4	0.2

@ mindful maths

$P(X = x)$

probability that

outcome of the random variable X...

... was the specific outcome x

Example: random variable X represents no. of heads when three coins are tossed.

Sample Space: HHH, HHT, HTT, TTT, TTH, THH, THT, HTH

Probability Distribution:

No. of heads	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Probability Function: $P(X = x) = \begin{cases} \frac{1}{8} & x = 0, 3 \\ \frac{3}{8} & x = 1, 2 \\ 0 & x = \text{otherwise} \end{cases}$

$$P(0 < x < 3) = \frac{3}{8} + \frac{3}{8} = \frac{6}{8}$$

$$E(X) = \sum x \cdot P(x)$$

↑ expected value

Scanned with CamScanner

$$E(X^2) = \sum x^2 \cdot p(x)$$

$$\text{Var}(X) = E(X^2) - E(X)^2$$

CODING

@ mindful maths

Let $y = 3x - 10$

$$\bar{y} = 3\bar{x} - 10$$

$$\sigma_y^2 = (3\sigma_x)^2$$

$$E(aX+b) = a\,E(X) + b$$

$$\text{Var}(aX+b) = a^2\,\text{Var}(X)$$

Scanned with CamScanner

6. DISCRETE RANDOM VARIABLE

Discrete Random Variable → must vary and be a result of an experiment

- **Only certain** numerical values can be taken
- **Value depends** on the outcome of a random event
- **Can take any** of a range of specific values

PROBABILITY DISTRIBUTION

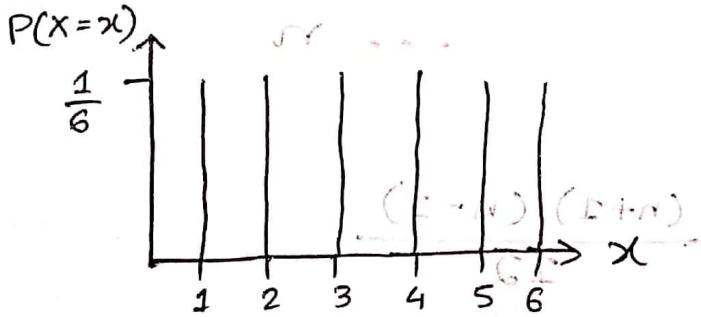
Fully describes the probability of any outcome in the sample space.

Let random variable X = 'score on a dice'

- probability function : $P(X = x) = \frac{1}{6}$, $x = 1, 2, 3 \ldots 6$
- table :

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- diagram:



Discrete uniform distribution

$$\sum P(X = x) = 1$$

$$P(X = x) = \begin{cases} \frac{1}{8} & x = 0, 3 \\ \frac{3}{8} & x = 1, 2 \end{cases}$$

Avishek

Scanned with CamScanner

Expected value

$$E(X) = \sum x \, P(X = x)$$

This value tells us the average of the values that could come in long run.

$$E(X^2) = \sum x^2 \, P(X = x)$$

$$\text{Var}(X) = E(X^2) - E(X)^2$$

$$E(aX + b) = a\,E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(aX + b) = a^2 \, \text{Var}(X)$$

FOR DISCRETE UNIFORM VARIABLE

Like – $1, 2, 3 \ldots n$

$$E(X) = \frac{n+1}{2}$$

$$\text{Var}(X) = \frac{(n+1)(n-1)}{12}$$

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Avishek

Scanned with CamScanner

7 THE NORMAL DISTRIBUTION

1) **Discrete** - Data that can take only certain values in a given range. (Example - No. of heads in 4 coin tosses)

2) **Discrete random variable** - Only takes values from a discrete scale.

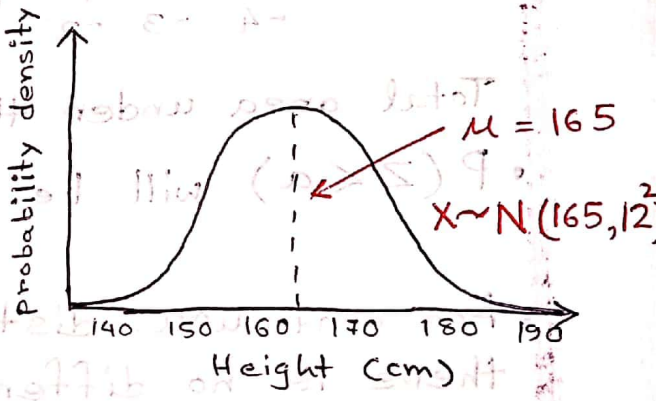
3) **Normal distribution** - Data that has been distributed or spread out in the shape of a bell curve.

- A continuous random variable has a continuous probability distribution.
- Area under a continuous probability distribution curve is equal to 1.

If X is a normally distributed variable, then you write:

$$X \sim N(\mu, \sigma^2)$$

mean ↑ variance ↑



☆ **The normal distribution:**

- has parameters μ and σ^2
- is symmetrical (mean = median = mode)
- has bell-shaped curve with asymptotes at each end
- has total area under the curve equal to 1
- has points of inflection at $\mu + \sigma$ and $\mu - \sigma$

(For a normally distributed variable, approximately 68%, 95% and 99.7% of data lies within ~~the~~ one, two and three standard deviations of the mean)

Avishek

Scanned with CamScanner

Using tables to find probabilities of standard normal distribution Z.

- Standard normal variable is usually denoted by **Z** and has a mean of 0 and a standard deviation of 1.

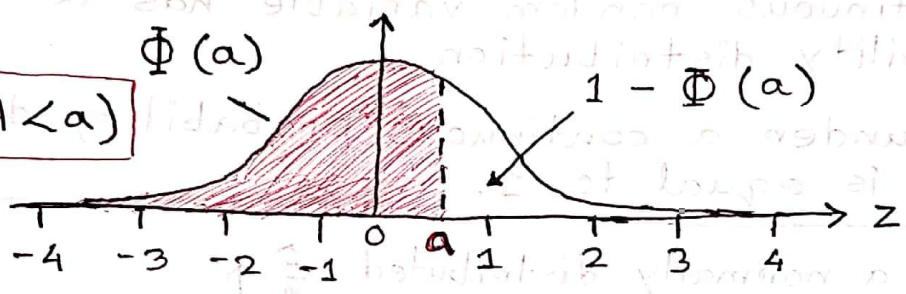
$\mu = 0, \sigma^2 = 1$

$$Z \sim N(0, 1^2)$$

$\sim$ means "is distributed"

normal

$\Phi(a) = P(A < a)$

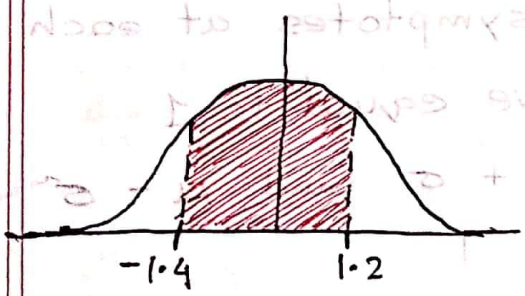


Total area under the curve = 1

- $P(Z < a)$ will be area under curve to the left of a

- For continuous distribution, such as normal distribution there is no difference between $P(Z < z)$ and $P(Z \leq z)$

EXAMPLE:



$$P(Z < 1.2) - P(Z < -1.4)$$
$$= P(Z < 1.2) - P(Z > 1.4)$$
$$= P(Z < 1.2) - [1 - P(Z < 1.4)]$$
$$= P(Z < 1.2) - 1 + P(Z < 1.4)$$
$$= 0.8849 - 1 + 0.9192$$
$$= 0.8041$$

Avishek

Scanned with CamScanner

Standard Normal Distribution

Normally distributed random variables can be standardised by coding the data so that it can be modelled as standard normal distribution.

Coding: $Z = \frac{X - \mu}{\sigma}$

$X \sim N(\mu, \sigma^2)$

1, 0

$P(Z > z)$

$= 1 - P(Z < z)$

$P(Z > -z)$

$= P(Z < z)$

$P(Z < -z)$

$= P(Z > z)$

$= 1 - P(Z < z)$

For $Z \sim N(0, 1^2)$

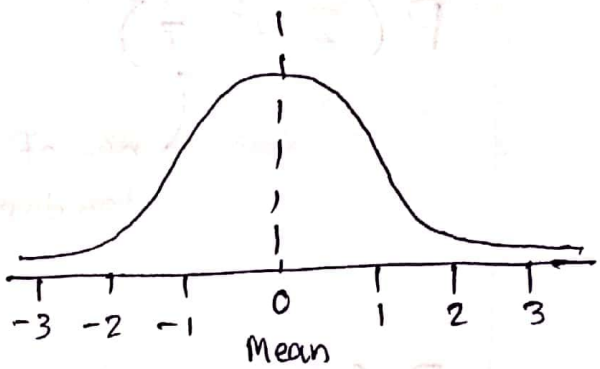
$P(Z < a) = \Phi(a)$

Avishek

Scanned with CamScanner

only for continuous variables

NORMAL DISTRIBUTION



- μ = mean
- area under curve = probability
- symmetrical graph

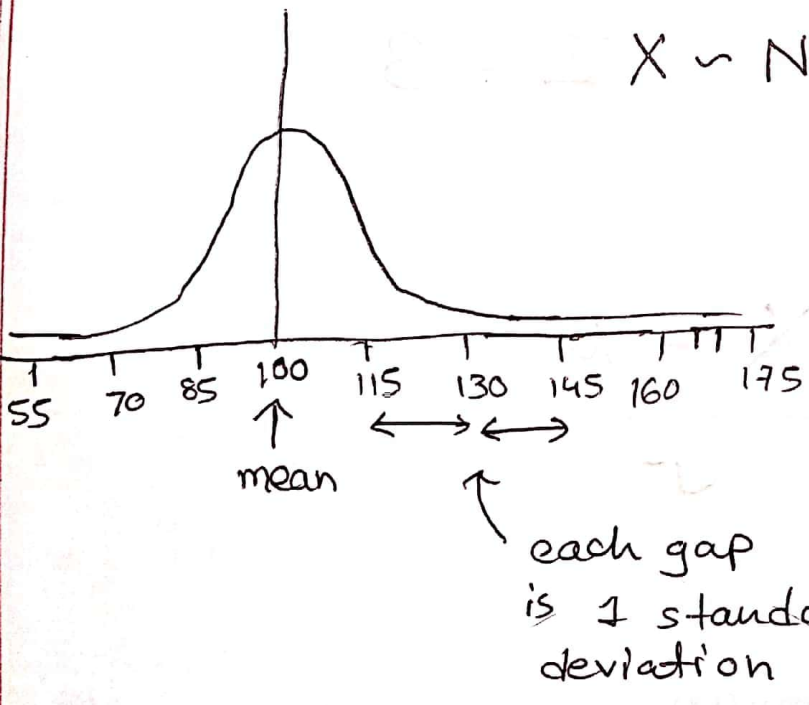
$$X \sim N(\mu, \sigma^2)$$

Random variable X — is distributed

Mindful Maths

Z is the no. of standard deviations a value is above the mean.

$$X \sim N(100, 15^2)$$



number	z
100	0
130	2
85	-1
165	4.3
62.5	-2.5

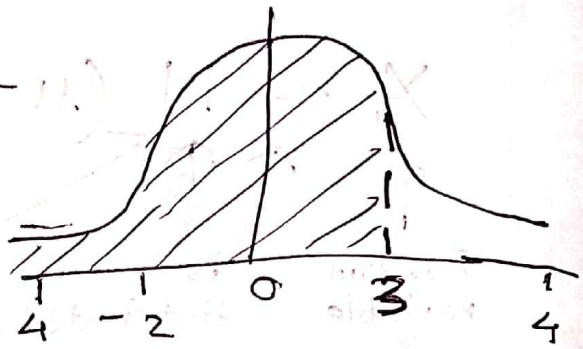
Scanned with CamScanner

$P(Z < z)$

↑ no. of standard deviation

If z value positive, then already half-way of graph.

$$P(Z < 3) = 1 - P(Z > 3)$$



$$P(Z < -z) = P(Z > z)$$

$$P(Z > -z) = P(Z < z)$$

$$P(Z < z) = 1 - P(z > z)$$

(X) world	(Z) world
$X = 145$	$Z = 3$

$\mu = 100$, $\sigma = 15$

Mindful Maths

$$Z = \frac{X - \mu}{\sigma}$$

$$\frac{\text{actual} - \text{mean}}{\text{standard deviation}}$$

$$P(a < X < b) = P(X < b) - P(X < a)$$

Scanned with CamScanner